IJCAI 2021 Workshop on Generalization in Planning

Learning Causal Models of Autonomous Agents using Interventions

Pulkit Verma, Siddharth Srivastava

Arizona State University

How Would End Users Assess Their AI Systems?

- How would a lay user determine whether an AI agent will be safe/reliable for a certain task?
- More challenging in settings where agent's internal code is not available (black-box).
- Can we get insights from how we assess humans in such situations?

Causal Models

- Understand the relationships among underlying causal mechanisms.
- Makes it easy to capture and predict the behavior of AI systems.
- Can be modeled using STRIPS-like models; precondition and effect maps to cause and effect.
- We use Halpern and Pearl's notion of actual cause†.

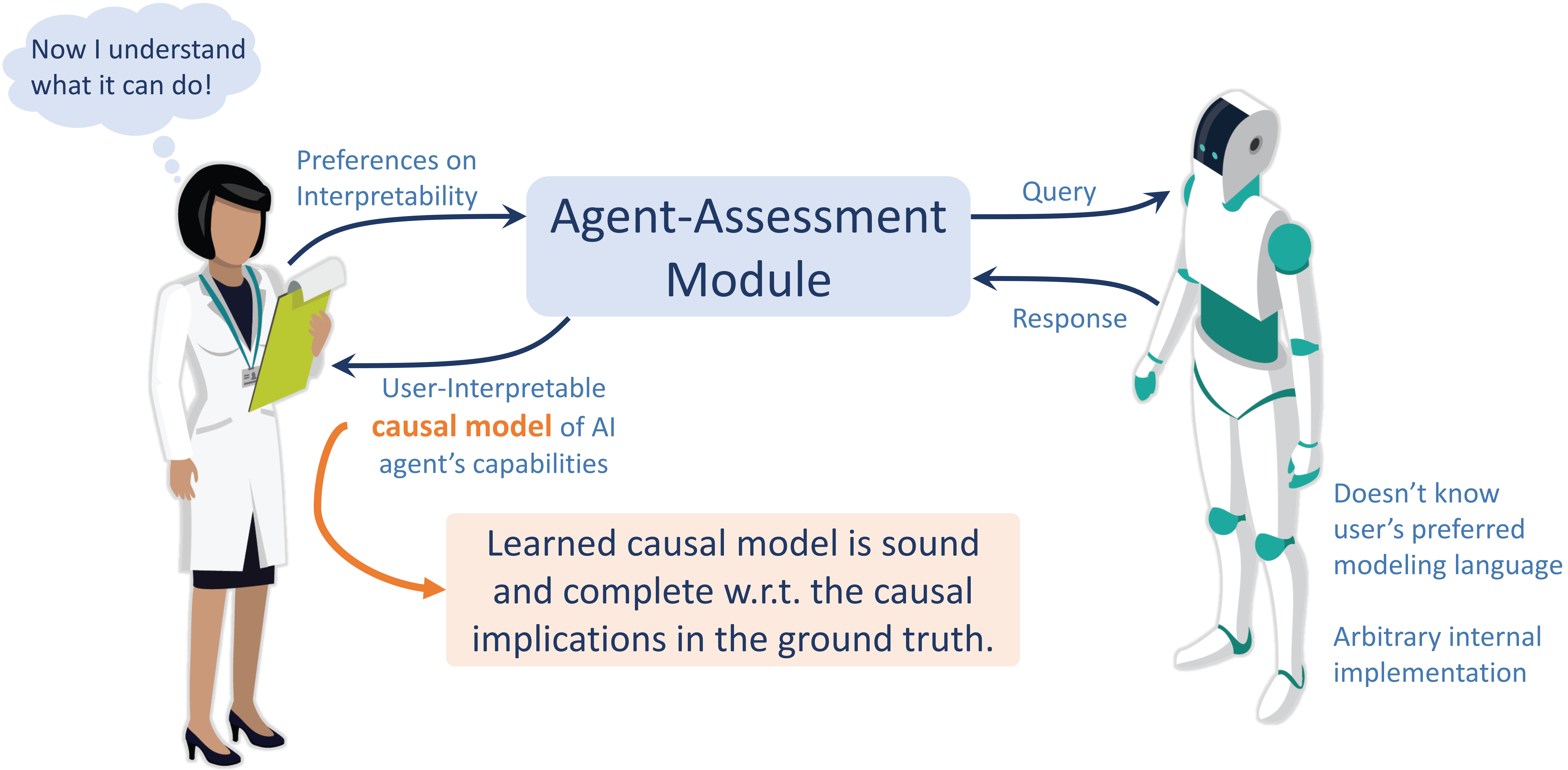
Refer to the paper for definitions of

- causal model†,
- soundness and completeness of a causal model w.r.t. the causal implications in the ground truth.

†Joseph Y. Halpern. Actual Causality. The MIT Press, 2016.

Related Work

- Several approaches have been developed for inferring interpretable models based on passively observed agent behavior.
  - E.g., ARMS – Yang et al. (AIJ 2007), LOCM - Cresswell et al. (ICAPS 2009), LOUGA - Kučera and Barták (KMAIS 2018), FAMA - Aineto et al. (AIJ 2019)
- Susceptible to unsafe model inference.
- Not guaranteed to be sound or complete w.r.t. the causal implications in the ground truth.



Now I understand what it can do!

Preferences on Interpretability

Agent-Assessment Module

Query

Response

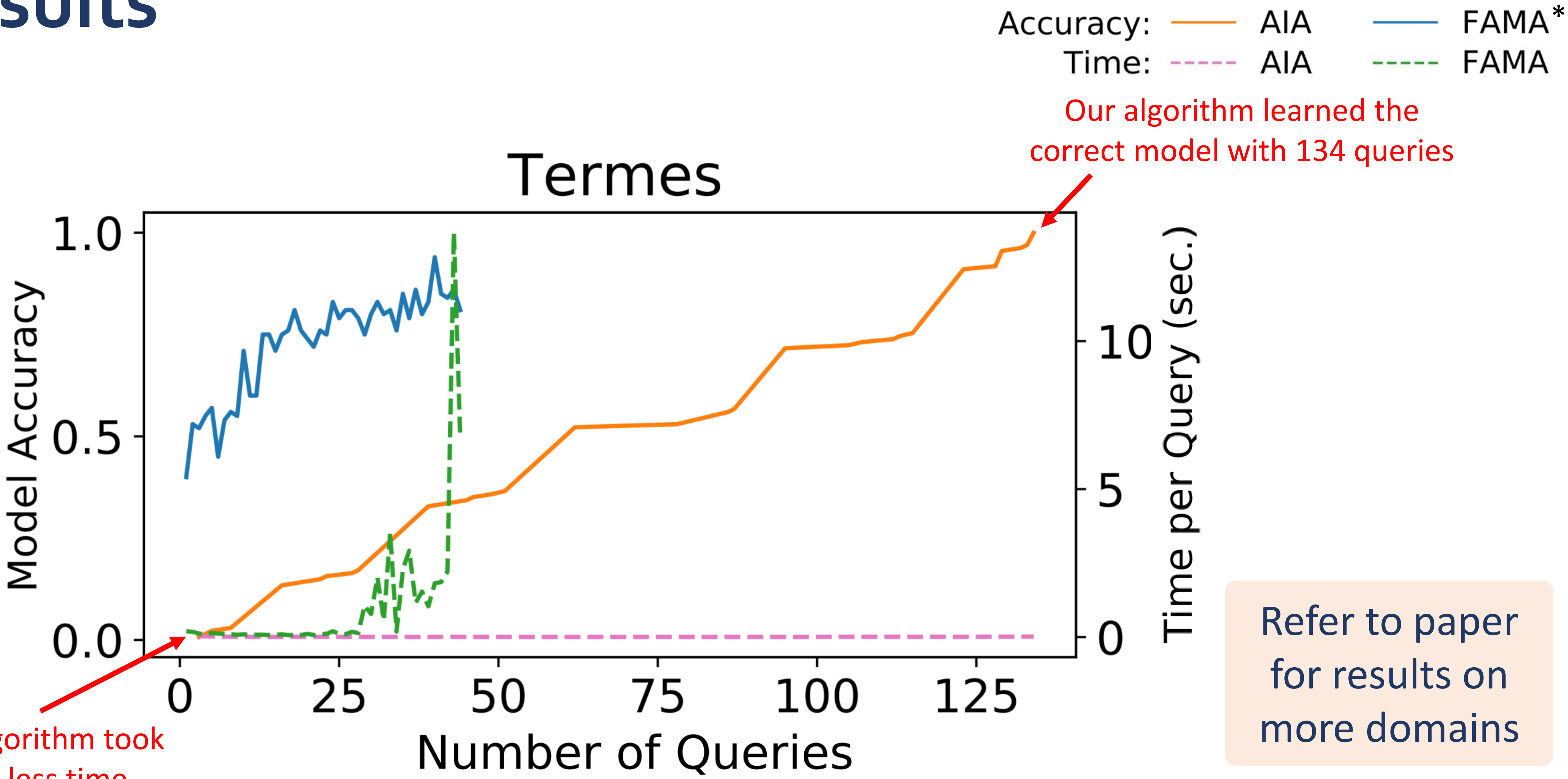
User-Interpretable **causal model** of AI agent's capabilities

Learned causal model is sound and complete w.r.t. the causal implications in the ground truth.

Doesn't know user's preferred modeling language

Arbitrary internal implementation

Results



Our algorithm learned the correct model with 134 queries

Our algorithm took very less time

Refer to paper for results on more domains

*Aineto, D.; Celorrio, S. J.; and Onaindia, E. 2019. Learning Action Models With Minimal Observability. AIJ, 275: 104–137.

Conclusions

The proposed approach:

- Efficiently learns causal model of an autonomous agent.
- Needs no prior knowledge of the agent model.
- Only requires an agent to have rudimentary query answering capabilities.
- Learns the model accurately with a small number of queries.

verma.pulkit@asu.edu

Visit the poster @ R3H, Red Montreal

AAAI'21 Paper

bit.ly/3p4cVRu

GenPlan'21 Paper

bit.ly/3eNcW9G